Open Source I/O Software for HPC: A Quick Tour

Rob Ross, Pete Beckman, Phil Carns, Jason Cope, Kevin Harms, Dries Kimpe, Kamil Iskra, Sam Lang, Rob Latham, Rusty Lusk, Seung Woo Son, Rajeev Thakur, Justin Wozniak

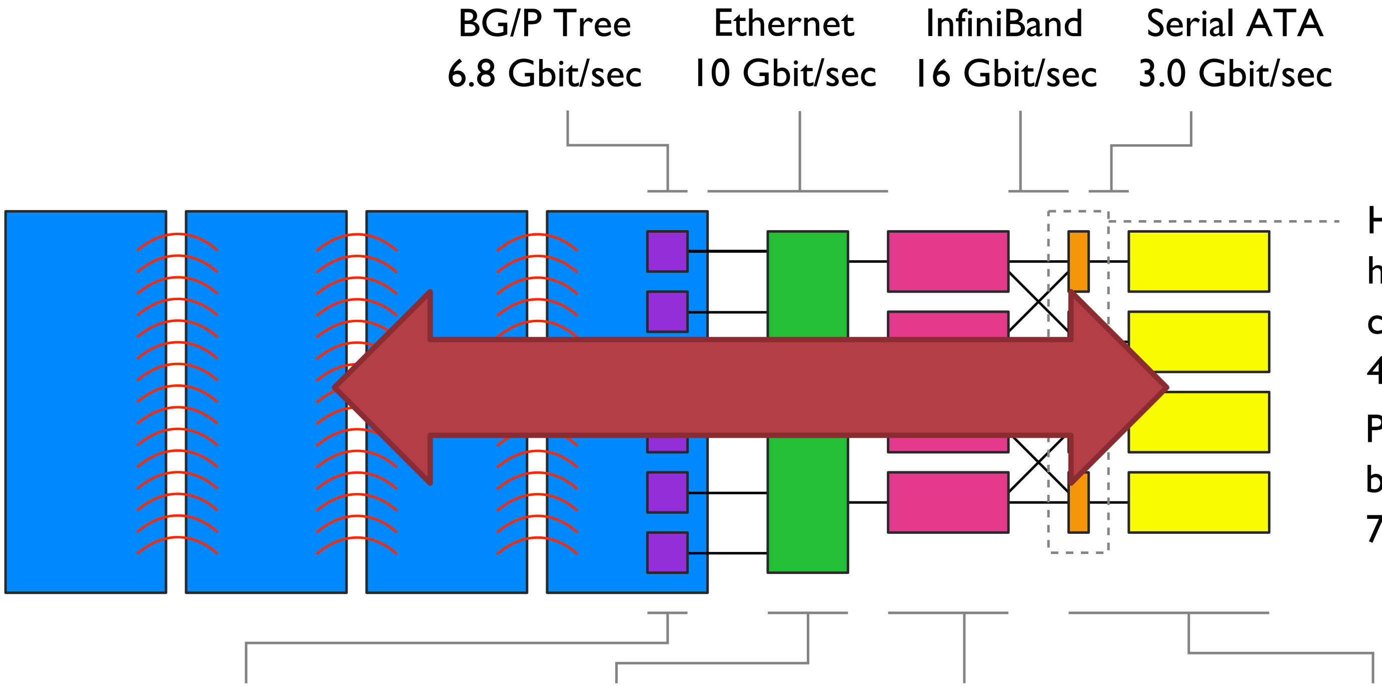
Mathematics and Computer Science Division

Argonne National Laboratory

rross@mcs.anl.gov

DOE Workshop for Industry Software Developers, March 31, 2011

Some Context…



BG/P Tree
6.8 Gbit/sec

Ethernet
10 Gbit/sec

InfiniBand
16 Gbit/sec

Serial ATA
3.0 Gbit/sec

HW bottleneck is here. Controllers can manage only 4.6 Gbyte/sec.

Peak I/O system bandwidth is 78.2 Gbyte/sec.

Gateway nodes run parallel file system client software and forward I/O operations from HPC clients.

640 Quad core PowerPC 450 nodes with 2 Gbytes of RAM each

Commodity network primarily carries storage traffic.

900+ port 10 Gigabit Ethernet Myricom switch complex

Storage nodes run parallel file system software and manage incoming FS traffic from gateway nodes.

136 two dual core Opteron servers with 8 Gbytes of RAM each

Enterprise storage controllers and large racks of disks are connected via InfiniBand or Fibre Channel.

17 DataDirect S2A9900 controller pairs with 480 1 Tbyte drives and 8 InfiniBand ports per pair

Architectural diagram of the 557 TFlop IBM Blue Gene/P system at the Argonne Leadership Computing Facility.

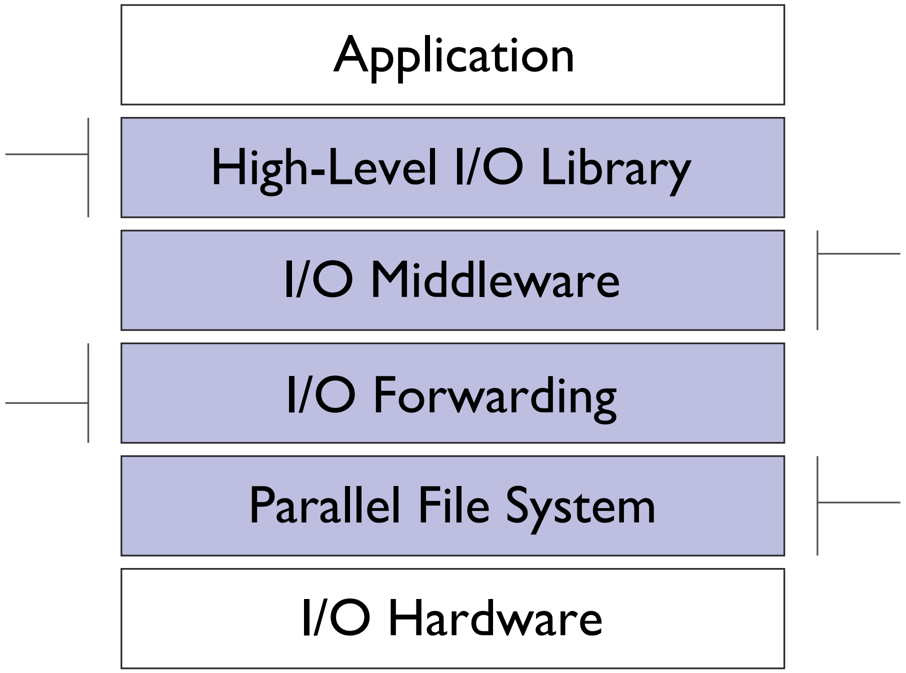
The HPC I/O Software Stack

High-Level I/O Library
maps application abstractions onto storage abstractions and provides data portability.

HDF5, Parallel netCDF, ADIOS

I/O Forwarding
bridges between app. tasks and storage system and provides aggregation for uncoordinated I/O.

IBM ciod



I/O Middleware
organizes accesses from many processes, especially those using collective I/O.

MPI-IO

Parallel File System
maintains logical space and provides efficient access to data.

PVFS, PanFS, GPFS, Lustre

Computational science applications are benefitting from a collection of additional software, built by the community, that is helping to extend the useful lifetime of the current HPC storage system approach. The Mathematics and Computer Science Division at Argonne develops and supports many of these components.

Widely Used I/O Software Developed at Argonne

Parallel Virtual File System (PVFS). A production-grade parallel file system for use in leadership computing systems and as basis for further PFS research.

- Collaboration with Clemson University.
- Sam Lang, Phil Carns, Rob Latham, Seung Woo Son.

ROMIO MPI-IO implementation. A portable, production-grade MPI-IO implementation for use in MPI applications on the widest possible variety of environments and as the basis for further research in MPI-IO.

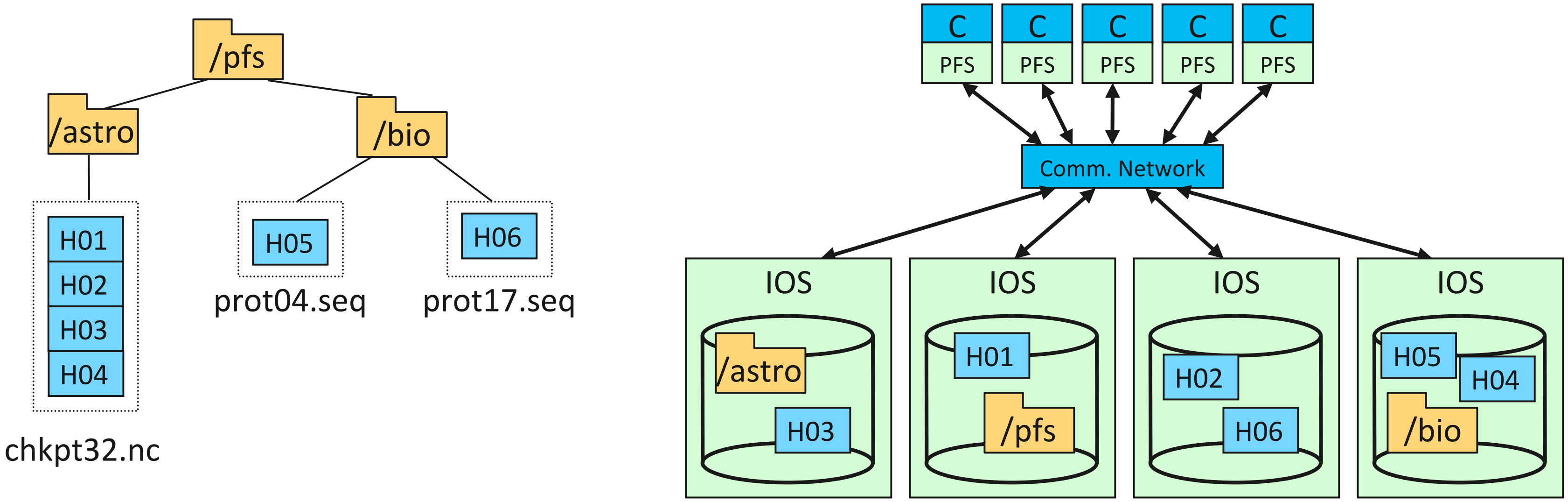
- Collaboration with Northwestern University.
- Rob Latham (Technical Lead), Dries Kimpe, Rob Ross, Rajeev Thakur.

Parallel netCDF (PnetCDF) high-level I/O library. An efficient interface for concurrent access to netCDF format datasets for highly parallel applications.

- Collaboration with Northwestern University.
- Rob Latham (Technical Lead), Dries Kimpe, Rajeev Thakur, Rob Ross.

The Parallel Virtual File System

The PVFS Parallel File System



An example parallel file system, with large astrophysics checkpoints distributed across multiple I/O servers (IOS) while small bioinformatics files are each stored on a single IOS.

- Work in collaboration with Clemson University
- Building block for HPC I/O systems
  - Present storage as a single, logical storage unit
  - Stripe files across disks and nodes for performance
  - Tolerate failures (in conjunction with other HW/SW)
- Linux kernel support, genesis as a Beowulf computing file system
- User interface is POSIX file I/O interface, with optional MPI-IO access

Industry use of PVFS: Acxiom Corporation

- "… the world's largest processor of consumer data …" – Fortune magazine
- Multi-national (US, UK, France, Australia, Japan)
- Houses data and runs analytics applications for financial and other large businesses
  - Compute and data intensive
  - 24/7 operation
  - Highly-available, redundant resources
  - 7000+ compute nodes deployed in widely distributed environment

ACXIOM®
REDISCOVER YOUR POTENTIAL™

ACXIOM in the news

Acxiom and NDCHealth Announce a Strategic Technology Infrastructure Agreement Aimed at Enabling 'Best-in-Class' Information Solutions to Healthcare Industry

ATLANTA, Ga., and LITTLE ROCK, Ark. – September 29, 2005 – NDCHealth® Corporation (NYSE: NDC) and Acxiom® Corporation (Nasdaq: ACXM) today announced a long-term strategic initiative whereby Acxiom will transform and manage the technology infrastructure of NDCHealth's Information Management business unit. The collaborative project will create a data factory for NDCHealth utilizing Acxiom's automated grid-enabled information management platform.

Acxiom® Expands Fraud Prevention, Risk Mitigation Capabilities with Acquisition of InsightAmerica

Acxiom® Launches Online Segmentation Solution
June 29, 2005

Acxiom Announces Success In Invitations-to-Apply Solution Testing, General Commercial Availability
June 29, 2005

Acxiom Again Listed Among Computerworld 100 Best Places to Work in Information Technology
June 27, 2005

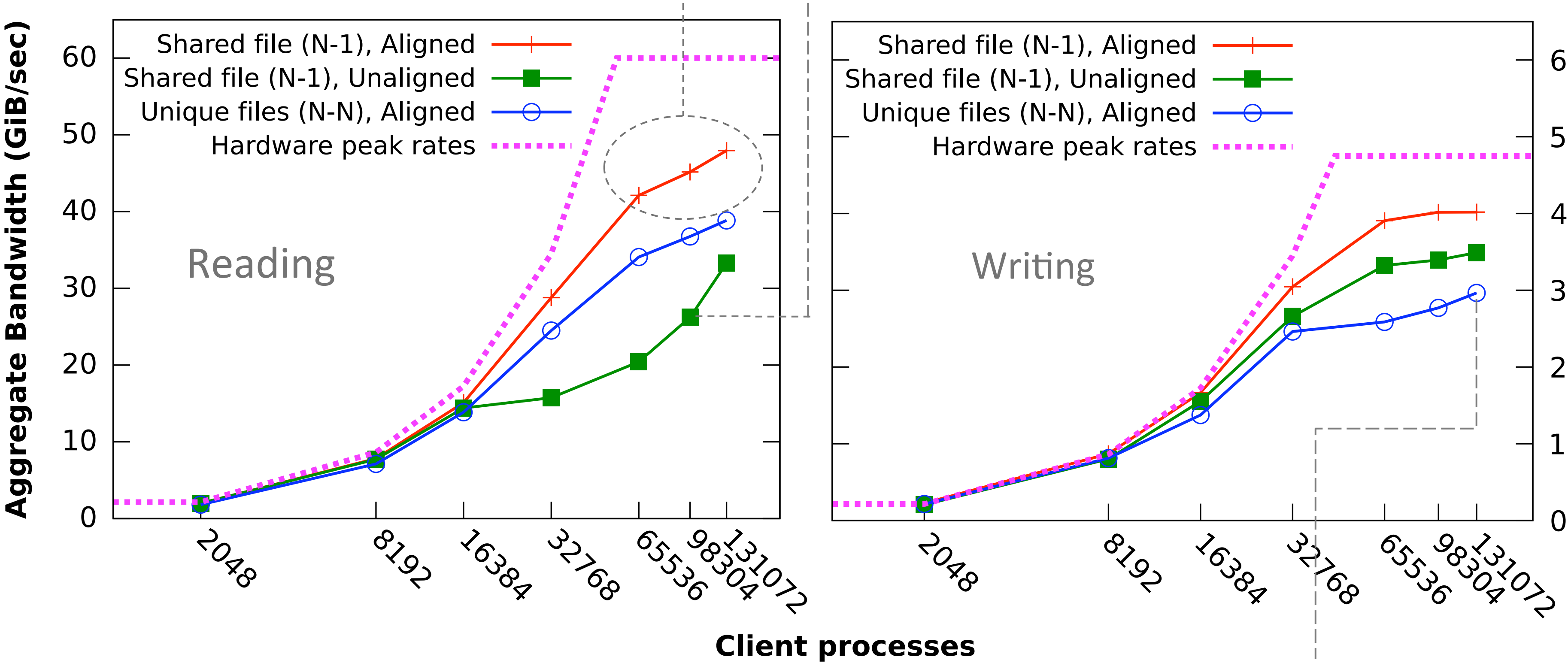
Acxiom's Customer Information Infrastructure Receives Computerworld Honors Program 21st Century Achievement Award
June 08, 2005

Hurricane Katrina - Red Cross Donations
Acxiom invites you to join us in contributing to the Hurricane Katrina relief effort.

PVFS Performance

Effective BW out of storage racks limited by scheduling and pipelining.

Unaligned requests are split, resulting in small (and odd-sized!) requests to servers



Unique file write allows hot spots in data placement, O(n) times more metadata updates for file size

- Used IOR
- Scalability testing uncovers drop from peak
- Clear differences in access patterns at higher process counts

PVFS Availability and Support

- Source available under LGPL (and GPL for kernel components) license

  http://www.pvfs.org
- Active mailing lists for users and developers, community support
- Clemson University has taken over technical lead in last 6 months
  - Their “OrangeFS” version has option for commercial support

    http://www.omnibond.com/orangefs/index.html
  - “OrangeFS” is PVFS, and their version will be the trunk in the near future

The ROMIO MPI-IO Library

The ROMIO MPI-IO Library

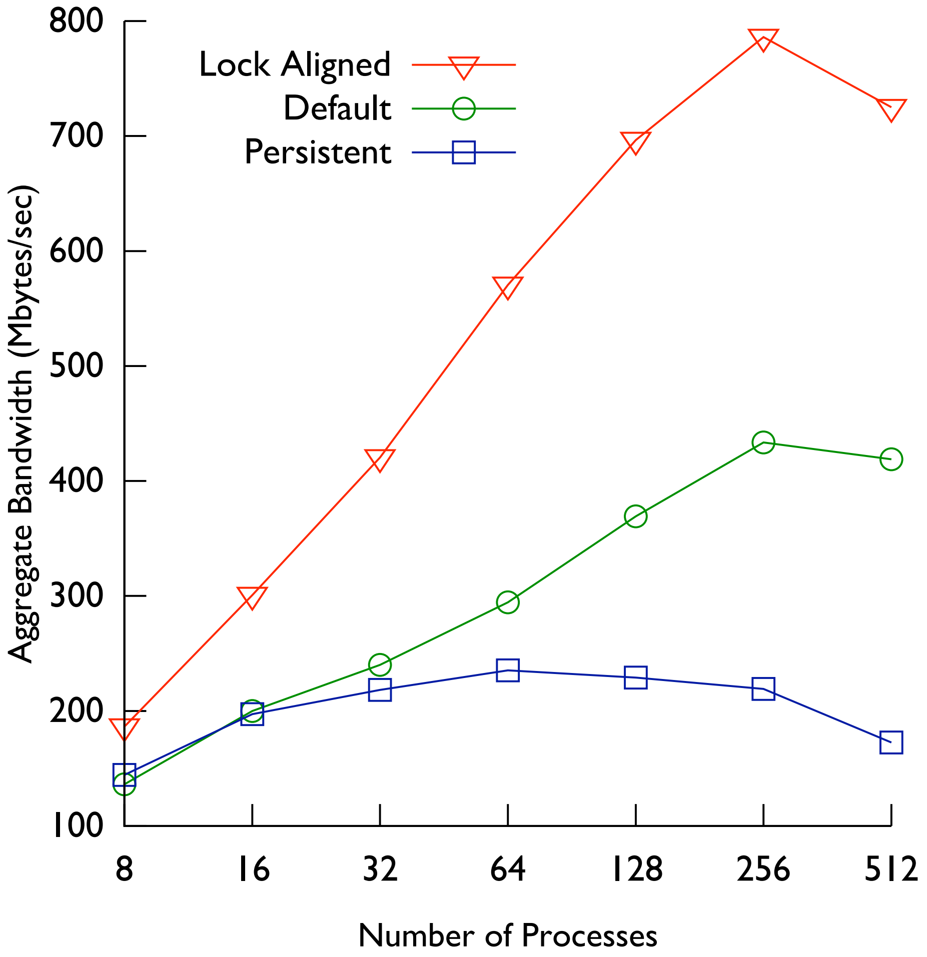
- Collaboration with Northwestern University and others
- ROMIO is a portable implementation of the I/O portion of the Message Passing Interface standard (MPI-IO)
  - Can be thought of as a transformation layer between an MPI application (or higher-level libraries) and the parallel file system
  - Goal of transformations is to improve performance by reducing number of operations to the file system, align accesses, etc.
  - Could be a place to manage on-node parallelism (hasn't been addressed yet)
- Origins in the days of Intel Paragons, etc., but adapted to newest systems
- Part of the MPICH2 MPI implementation
- Integrated into products from Cray, IBM, SGI, Intel, and others
  - …and many of these vendors contribute back to the source tree!

MPI-IO Optimizations at Work

Graph shows aggregate bandwidth for simulated checkpointing from a combustion application (S3D) on the Mercury system at NCSA, using GPFS and collective MPI-IO write.

Significant gains are had by aligning accesses with the granularity of locks. Interestingly, the "persistent" optimization is most effective on Lustre volumes because of differences in how locks are managed.

These types of improvements make concurrently writing feasible at scale, which is critical for checkpointing.



W.-K. Liao and A. Choudhary, "Dynamically adapting file domain partitioning methods for collective I/O based on underlying parallel file system locking protocols," SC2008.

ROMIO Availability and Support

- Source available under BSD-like license (approved by IBM, Intel, Microsoft, etc.)
  - Best obtained as part of MPICH2

    http://www.mcs.anl.gov/mpich2
- Mailing list for users and support, but typically support is provided by the system vendor (e.g., IBM)

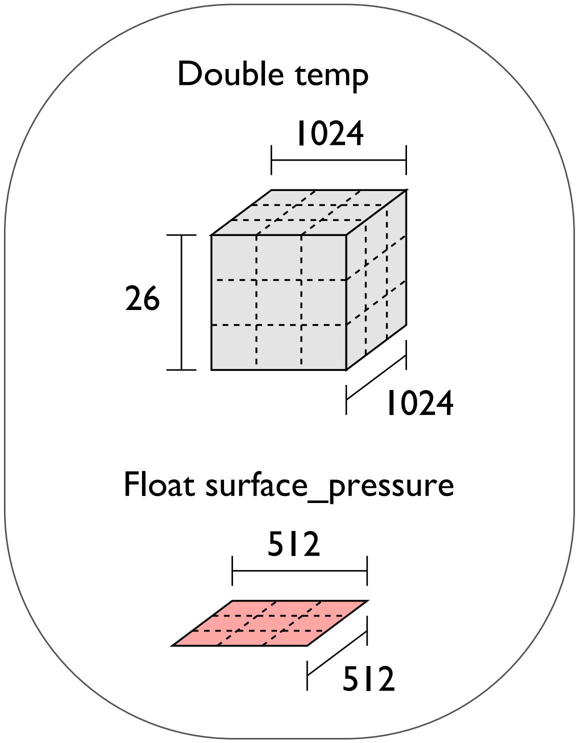
Parallel netCDF

The Parallel netCDF (PnetCDF) High-Level I/O Library

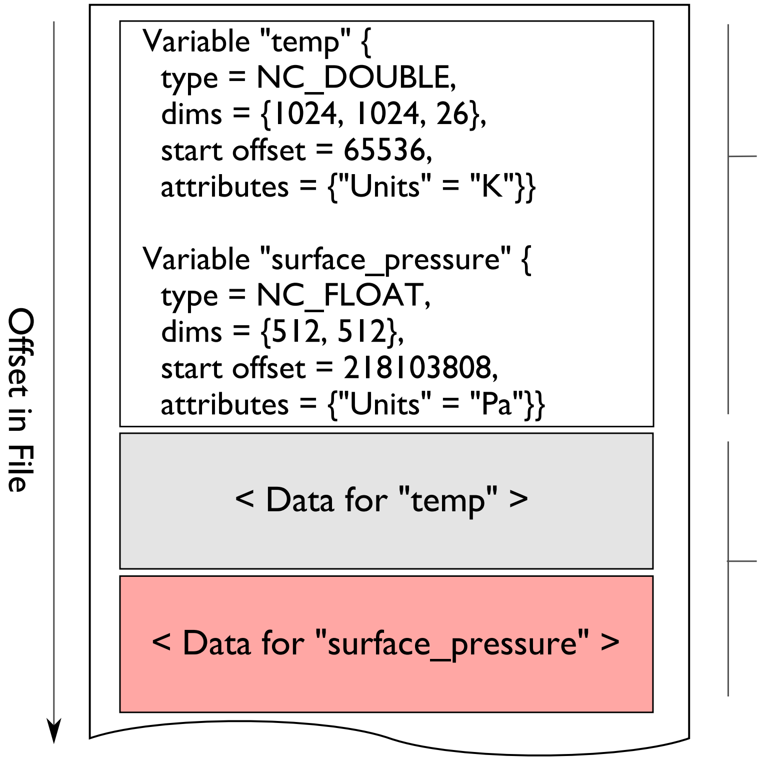
- Work in collaboration with Northwestern University
- Based on original "Network Common Data Format" (netCDF) work from Unidata
  - Derived from their source code
- Data Model:
  - Collection of variables in single file
  - Typed, multidimensional array variables
  - Attributes on file and variables
- Retains the file format used by many climate and weather codes, as well as other HPC applications
- Newest of these three projects
  - Developed in early 2000s as part of SciDAC-1 effort
  - Addressed critical needs in climate science

High-level I/O Libraries and Productivity

Application Data Structures



netCDF File "checkpoint07.nc"



```
Variable "temp" {
  type = NC_DOUBLE,
  dims = {1024, 1024, 26},
  start offset = 65536,
  attributes = {"Units" = "K"}}

Variable "surface_pressure" {
  type = NC_FLOAT,
  dims = {512, 512},
  start offset = 218103808,
  attributes = {"Units" = "Pa"}}
```

< Data for "temp" >

< Data for "surface_pressure" >

netCDF header describes the contents of the file: typed, multi-dimensional variables and attributes on variables or the dataset itself.

Data for variables is stored in contiguous blocks, encoded in a portable binary format according to the variable's type.

High-level I/O libraries help application scientists map their data into storage abstractions. Secondarily, they (hopefully) perform this mapping in an efficient manner, such as padding data to appropriate boundaries.

PnetCDF Availability and Support

- Source available under BSD-like license

  http://www.mcs.anl.gov/projects/parallel-netcdf
- Active support mailing list
- Portable across Linux, Cray, IBM, etc. platforms
  - Relies on MPI-IO for I/O accesses

In Summary…

- Active community effort in developing and supporting tools for I/O in HPC systems
- Argonne is a participant in three such efforts:
  - **Parallel Virtual File System (PVFS)**. A production-grade parallel file system for use in leadership computing systems and as basis for further PFS research.
  - **ROMIO MPI-IO implementation**. A portable, production-grade MPI-IO implementation for use in MPI applications on the widest possible variety of environments and as the basis for further research in MPI-IO.
  - **Parallel netCDF (PnetCDF) high-level I/O library**. An efficient interface for concurrent access to netCDF format datasets for highly parallel applications.
- All these products are used in production at large scale today, freely available, with limited support infrastructure
- Happy to see adoption by industry partners, especially if they contribute bug fixes back to us

Acknowledgments

The work presented was supported in full by the DOE Office of Advanced Scientific Computing Research (ASCR) under contract DE-AC02-06CH11357.